2

AFHRL-TP-90-14

AIR FORCE

HUMAN RESOURCES

LABORATORY

GENERALIZABILITY OF WALK-THROUGH PERFORMANCE TESTS, JOB PROFICIENCY RATINGS, AND JOB KNOWLEDGE TESTS ACROSS EIGHT AIR FORCE SPECIALTIES

Kurt Kraiger

Department of Psychology
University of Colorado at Denver
1200 Larimer Street
Denver, Colorado 80204

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

July 1990
Interim Technical Paper for Period October 1987 - February 1990

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

MARK TEACHOUT
Contract Monitor

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Colonel, USAF
Chief, Training Systems Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
July 1990

3. REPORT TYPE AND DATES COVERED
Interim Paper - October 1987 to February 1990

4. TITLE AND SUBTITLE
Generalizability of Walk-Through Performance Tests, Job Proficiency Ratings, and Job Knowledge Tests Across Eight Air Force Specialties

5. FUNDING NUMBERS

C	-	F41689-86-D-0052
PE	-	62205F
PR	-	1121
TA	-	13
WU	-	01

6. AUTHOR(S)
Kurt Kraiger

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Universal Energy Systems
8961 Tesoro Drive, Suite 600
San Antonio, Texas 78217

P.C.

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Training Systems Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFHRL-TP-90-14

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Generalizability theory was used to assess the psychometric quality of Walk-Through Performance Tests (WTPTs) and Job Proficiency ratings in eight Air Force occupational specialties. In addition, generalizability theory was used to determine whether proficiency ratings and job knowledge test scores were substitutable for the WTPTs. The results showed that both the the WTPT scores and ratings within rating sources were generalizable (reliable), but that ratings were not generalizable over rating sources, and neither ratings nor job knowledge test scores ranked incumbents similarly to the WTPT. KEYWORDS

14. SUBJECT TERMS
~~D-Study~~
~~G-Study~~
~~generalizability~~ theory
~~job performance measurement~~
reliability

15. NUMBER OF PAGES
70

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

AFHRL Technical Paper 90-14 July 1990

GENERALIZABILITY OF WALK-THROUGH PERFORMANCE TESTS, JOB PROFICIENCY RATINGS, AND JOB KNOWLEDGE TESTS ACROSS EIGHT AIR FORCE SPECIALTIES

Kurt Kraiger

Department of Psychology
University of Colorado at Denver
1200 Larimer Street
Denver, Colorado 80204

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

James B. Bushman, Major, USAF
Chief, Training Assessment Branch
Training Systems Division

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This paper summarizes the application of generalizability (G) theory to the Air Force Job Performance (JPM) project. Generalizability analyses were applied using three different sets of performance measures for eight occupational specialties. More specifically, G theory was used to assess the dependability of performance scores over different performance rating conditions (i.e., rating sources, rating forms, or rating dimensions), different Walk-Through Performance Test (WTPT) conditions (hands-on vs. interview assessment, different job tasks, or different steps within tasks), and over different general measurement techniques (ratings, WTPTs, or job knowledge tests). Ratings were found to be generalizable within rating sources, and WTPT scores were found to be generalizable over methods, tasks, and steps. Ratings were not generalizable over rating sources, and neither ratings nor job knowledge tests were substitutable for WTPT scores. Results of these analyses were consistent over occupational specialties, particularly for the rating variables.

PREFACE

The Air Force Job Performance Measurement (JPM) project is a large-scale, multi-faceted effort to assess individual job proficiency. Within the specialties examined herein, incumbents are assessed via Walk-Through Performance Tests (WTPT), job proficiency ratings, and (for some specialties), job knowledge tests.

A critical issue concerns the psychometric quality of these various measures. The present study supports the JPM project by assessing the psychometric quality of both the WTPT and rating methods, and by examining the extent to which ratings and the job knowledge tests are substitutable for the WTPT. In addition, the results of these analyses are compared over specialties to determine the extent to which judgments of measurement quality based on data collection to date are warranted. These issues are addressed primarily through the application of generalizability (G) theory. G theory identifies whether scores assigned to individuals are dependable (or consistent) over conditions of measurement. For the rating data, the relevant conditions of measurement were rater sources, rating forms, and items or dimensions within particular forms. For the WTPT, relevant conditions of interest were assessment method (hands-on vs. interview), tasks, and steps or items within tasks. For the substitutability issue, a third generalizability design was constructed with performance measures (WTPT scores, ratings, and job knowledge test scores) and tasks as the conditions of interest. Finally, for both the WTPT and rating measures, a subset of generalizability analyses known as D studies was employed to investigate the dependability of these measures under specific measurement conditions (e.g., a single rating source or a single WTPT method).

The author greatly acknowledges the efforts of Mr. Mark Teachout of the Air Force Human Resources Laboratory toward

the completion of this paper. Mark aided the completion of this paper by sharing his knowledge of the JPM project and by his timely review of earlier drafts of this manuscript.

TABLE OF CONTENTS

Page

I. INTRODUCTION..1

Job Performance Measures..........................3

Current G-Theory Investigation....................4

II. METHOD..5

Sample..5

Rating Facets of Generalization...................5

Facets for WTPT Data................................7

Facets for Substitutability Design................8

D Study Analyses....................................9

III. RESULTS..10

Ratings Design......................................10

Within Source Analyses.............................24

G Study Results, WTPT Data.........................27

D Study Results, WTPT Data.........................29

G and D Study Results, Substitutability Design....30

IV. DISCUSSION..32

Psychometric Quality of Performance Ratings.......33

Psychometric Quality of WTPT Scores...............35

Other Measures as Surrogates for WTPT Scores......37

Recommendations.....................................38

REFERENCES..40

APPENDIX A..43

LIST OF FIGURES

FIGURE Page

1 G Coefficients for Performance Rating Data for Eight Occupational Specialties................23

2 G Coefficients Within Rating Sources for Eight Occupational Specialties.........................26

3 G Coefficients for WTPT Score for Eight Occupational Specialties..........................31

LIST OF TABLES

TABLE Page

1 Descriptive Statistics for Rating Variables, for Jet Engine Mechanics..........................12

2 Descriptive Statistics for Rating Variables, for Avionic Communications Specialists............13

3 Descriptive Statistics for Rating Variables, for Air Traffic Control Operators.................14

4 Descriptive Statistics for Rating Variables, for Information Systems Radio Operators...........15

5 Descriptive Statistics for Rating Variables, for Aircrew Life Support..........................16

6 Descriptive Statistics for Rating Variables, for Personnel Specialist..........................17

7 Descriptive Statistics for Rating Variables, for Equipment Laboratory Specialist...............18

8 Descriptive Statistics for Rating Variables, for Aerospace Ground Equipment...................19

9 Estimateu Variance Components for G Study of Rating Variables with Three Forms.................21

10 G and D Study Results for Within Source Analyses..25

11 Estimated Variance Components for G Study of WTPT Variables with Tasks, Methods Crossed 28

12 Estimated Variance Components for G Study of WTPT Variables With Tasks Nested in Methods 29

13 G and D Study Results for Substitutability Design using Supervisory Ratings, WTPT Scores, and Job Knowledge Test Scores.............................33

A-1 Estimated Variance Components for Aircrew Life Support, Three-Form Analysis.....................43

A-2 Estimated Variance Components for Personnel Specialist, Three-Form Analysis..................44

A-3 Estimated Variance Components for Precision Measurement Equipment Laboratory Specialist, Three-Form Analysis.............................45

A-4 Estimated Variance Components for Aeropace Ground Equipment, Three-Form Analysis.... 46

A-5 Simulated D Study Results of Rating Analysis for Aircrew Life Support..............................47

A-6 Simulated D Study Results of Rating Analysis for Personnel Specialist..............................48

A-7 Simulated D Study Results of Rating Analysis for Precision Measurement Equipment Laboratory Specialist......................................49

A-8 Simulated D Study Results of Rating Analysis for Aerospace Ground Equipment.......................50

A-9 G Study Results for Crossed Design Analysis of WTPT Scores, Personnel Specialist................51

A-10 G Study Results for Crossed Design Analysis of WTPT Scores, Aerospace Ground Equipment...........52

A-11 G Study Results for Nested Design Analysis of WTPT Scores, Aircrew Life Support................53

A-12 G Study Results for Nested Design Analysis of WTPT Scores, Personnel Specialist................54

A-13 G Study Results for Nested Design Analysis of WTPT Scores, Precision Measurement Equipment Laboratory Specialist............................55

A-14 G Study Results for Nested Design Analysis of WTPT Scores, Aerospace Ground Equipment...........56

A-15 Simulated D Study Results for WTPT Analysis of Personnel Specialist.............................57

A-16 Simulated D Study Results for WTPT Analysis of Aerospace Ground Equipment......................57

GENERALIZABILITY OF WALK-THROUGH PERFORMANCE TESTS, JOB PROFICIENCY RATINGS, AND JOB KNOWLEDGE TESTS ACROSS EIGHT AIR FORCE SPECIALTIES

I. INTRODUCTION

The major goal of the Air Force performance measurement project is to provide the necessary data to establish valid linkages between enlistment standards and job performance. To this end, the staff for the Air Force Job Performance Measurement (JPM) System has applied Walk-Through Performance Tests (WTPT) and proficiency rating methodologies to data collection in four specialties, and WTPT, proficiency ratings, and job knowledge tests to data collection in an additional four specialties. The objective of the present paper is to support the development of these measures by: using Generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to assess the psychometric quality of both the WTPT and the performance ratings, examining the extent to which either proficiency ratings or job knowledge tests are substitutable for the WTPTs, and then comparing results of these analyses across multiple occupational specialties. Stated in G theory terminology, the purpose of the present investigation is to determine whether the evaluation systems yield dependable scores over conditions of measurement and whether measured incumbent performance levels are dependable over various evaluation methods.

Generalizability theory was developed by Cronbach and his associates (Cronbach et al., 1972) as an alternative to classical test theory. Whereas classical theory permits only univariate investigations of the effects of measurement error on reliability, G theory permits multifaceted analysis of the dependability of scores over a variety of measurement conditions. Recent detailed discussions and reviews of G theory may be found in Kraiger (1989) and Shavelson, Webb, and Rowley (1989).

G theory answers the question, "Does it matter if...?" That is, generalizability analyses can determine the relative variance in scores which can be attributable to various conditions of measurement. If variance over conditions is low, overall scores are said to generalize over the conditions of measurement. More informally, low variability over conditions implies that it "doesn't matter if" the measure is operationalized in different ways. Said yet another way, generalizability analyses indicate the degree to which scores based on a limited opportunity for observation (e.g., a work sample on a single occasion) are dependable over a considerably broader sample of possible observations (e.g., other tasks, occasions, etc.)

In any generalizability study, the researcher must first identify any factors of interest which could affect the measurement process. The researcher then must specify a particular range of levels for each factor. In G theory terminology, factors of measurement are called facets and levels of the facet are called conditions. An individual's average score over all combinations of conditions is said to be that person's universe score. Generalizability (G) studies are conducted to estimate the contribution to total score variance of the each facet and their interactions. Variance components are estimated for each effect and represent estimated variance about universe scores for average single observations, e.g., an average person evaluated by a single rater on a single occasion. In addition, a summary generalizability coefficient could be computed from individual variance components. This coefficient is analogous to a reliability coefficient in classical test theory and represents the proportion of observed score variance which is attributable to individual differences in the attribute being assessed. However, interpretations of individual variance components are often more enlightening since these reflect contributions to error variance by particular aspects of the measurement system (Brennan & Kane, 1979) and may be interpreted as evidence of construct validity (Kraiger & Teachout, 1990).

While G studies are useful for identifying the general characteristics of a measuring device, they may be misleading for describing the psychometric quality of an instrument under actual or intended circumstances. This is because G study variance components are estimated for single items or single administrations, even though organizations often use multiple operationalizations of constructs (e.g., multiple-item scales). Thus, a researcher may wish to perform a decision (D) study to assess the specific characteristics of a measurement instrument in a particular decision-making context. Similar to the Spearman-Brown prophecy formula in classical test theory, D studies allow a researcher to forecast resulting variance components and generalizability coefficients under different sets of measurement conditions. While the Spearman-Brown formula permits estimation when only a single parameter (typically items) is varied, D studies allow estimation of estimated effects when multiple facets are simultaneously varied. For example, generalizability coefficients can be estimated when ratings are averaged over three raters on a single occasion or two raters on two occasions. D study results often are of the most interest to decision-makers since they reflect realistic or intended measurement conditions.

Job Performance Measures

The Air Force JPM project assesses incumbent work proficiency via three mechanisms: WTPTs, job performance ratings, and job knowledge tests. The benchmark method is the WTPT; it includes both observation of actual hands-on performance and incumbent interviews. The WTPT hands-on format requires job incumbents to perform a series of job tasks under the careful observation of a highly-trained test administrator. The interview format requires incumbents to describe in detail the steps they would perform to accomplish various job tasks. In addition, airmen are assessed on four different rating forms by three different sources: Incumbents themselves, one to three peers, and an immediate supervisor. Each rating form assesses individual proficiency via a 5-point rating scale. These assessment methods are described in more detail in Hedge and Teachout (1986). Finally, incumbents in four specialties are also assessed via job knowledge tests. These tests require incumbents to answer multiple-choice questions regarding critical on-the-job tasks. Additional details on the job knowledge test are provided in Bentley, Ringenbach, and Augustin (1989).

Current G-Theory Investigation

Generalizability theory was used to address issues involving the dependability of ratings and WTPT data, and the substitutability of ratings and job knowledge tests for WTPT scores.

Generalizability of Rating Data. The first area of inquiry was the generalizability of performance ratings over different conditions of measurement. This investigation sought to extend the findings of Kraiger (1989; 1990; Kraiger & Teachout, 1987, 1990) to a total of eight Air Force specialties. Facets of interests were rating sources (incumbents, peers, and supervisors), forms (task-level, dimensional, global, and Air Force-side), and items (or scales/dimensions) nested within forms.

Generalizability of WTPT Data. The second area of interest was the generalizability of the WTPT scores. For each specialty, incumbents are assessed using both the hands-on and interview formats. This investigation sought to extend the findings of Kraiger (1990) to a total of eight AF specialties. Facets of interest were methods (hands-on vs. interview), the number of tasks assessed by either method, and the number of items or steps comprising individual tasks.

Substitutability Design. A third research question was whether performance ratings or performance ratings and job knowledge tests could be considered acceptable surrogates for the WTPT. In this design, assessment method (ratings, job knowledge tests and WTPT) and tasks were considered the primary facets. Separate analyses were performed for all rating sources combined, as well as each source individually.

A final research issue was the extent to which results of the research questions described above were consistent over specialties. G theory was not used to address this issue, but instead the results of the G studies in each specialty were compared and analyzed rationally.

II. METHOD

Sample

Proficiency ratings were collected from first-term airmen in eight different specialties. The specialties and their respective sample sizes were: Jet Engine Mechanic (AFS426x2), *n*=255; Air Traffic Controller Operator (AFS272x0), *n*=172; Avionic Communications Specialist (AFS328x0), *n*=98; Information Systems Radio Operator (AFS492x1, *n*=156; Aircrew Life Support (AFS122x0), *n*=216; Personnel Specialist (AFS732x0), *n*=218; Precision Measurement Equipment Laboratory Specialist (AFS324x0), *n*=138; and Aerospace Ground Equipment (AFS423x5), *n*=264. For all eight specialties, incumbent performance was measured by the WTPT and proficiency ratings. The WTPT was administered on their job site and required them to either perform or describe how they would perform the sampleu tasks. Their performance was observed by a carefully trained observer who recorded whether or not they executed (or described) the correct steps to accomplish the task. In addition, incumbents were rated on each of four rating forms by themselves, one or more peers, and their immediate supervisor. Finally, for the latter four specialties, incumbents also completed a job knowledge test, consisting of multiple-choice questions designed to assess an understanding of the tasks completed on the WTPT or rated on the forms. The generalizability of these measures was assessed through the analyses described below.

Rating Facets of Generalization

For the investigations of the performance rating data, there were three facets of generalization: Rating forms, specific items or scales included on each form, and rating sources. Items were nested with forms, and both were crossed with sources and ratees, yielding 11 distinct sources of variance.

Complete details concerning each facet are given in Kraiger (1989). The first facet of interest was rating sources, with incumbents, peers, supervisors as the conditions of the facet. The sources can be considered random samples of a larger universe of possible sources which could be used to assess ratee performance. When airmen were rated by more than one peer, only a single randomly-selected rating was used in order to balance the design. The second facet was rating forms, with task-level, dimensional, global, and Air Force-wide forms as the conditions of the facet. These can be considered random samples of a larger universe of possible forms which could be used to assess ratee performance. The final facet was the individual items, dimensions, or scales which comprise each form. Again, the items on any one form can be considered a random sample of possible items which could constitute that form. Items were nested within forms because individual items or scales vary from form to form.

As in Kraiger (1989; 1990), there was a computational problem with the items facet. This facet was unbalanced since the number of items on a form can range from two (on the global form) to over 30 (on the task-level form). Unbalanced facets may produce biased mean square estimates, which in turn are used to compute variance components (Searle, 1971). To compensate, analyses were run with two randomly selected items from all four forms and with *x* number of items from all forms except the shorter, 2-item global form, where *x* was the number of items on the dimensional form (the next shortest form). As in Kraiger (1989; 1990), results from both analyses were similar, and yielded comparable conclusions regarding the generalizability of ratings. For the sake of brevity, only the results of the three-form analyses are presented, as these contain less sampling error than the four-form analyses.

Facets for WTPT Data

For G study investigations of the WTPT, there were three facets of interest. The first facet was the method of assessment, with the hands-on and interview components as the conditions of the facet. The second facet was the tasks that were measured by both the hands-on and interview components. Typically, a WTPT consisted of 20-25 tasks. For each specialty, these tasks can be considered random samples of a larger possible universe of tasks which could comprise the WTPT. For purposes of analysis, there were two possible generalizability designs investigating variance due to tasks. For each specialty, there were three types of tasks included in the WTPT: Tasks common to both the hands-on and interview components, tasks unique to the hands-on component, and tasks unique to the interview component. Thus, common tasks were assessed by both methods, while unique tasks were assessed by one WTPT method but not the other. One analysis (the "crossed design") included only the common tasks and treated tasks as crossed with methods, since each task is assessed by each method and each method includes all tasks. A second analysis (the "nested design") included the unique tasks and treated tasks as nested within methods since tasks differed across methods of the WTPT. However, to increase the number of task conditions analyzed (and reduce sampling error in the entire design), analyses were conducted with the common tasks considered nested along with the unique tasks. That is, nested within a method might be eight unique tasks and six common tasks, even though these common tasks were not really nested. Results of these analyses were very similar to results from analyses using only unique tasks, but with smaller sampling error in the estimates of variance components. For some specialties, there were uneven numbers of tasks across the two methods. To balance the design, one or two unique tasks were randomly selected and discarded.

The final facet of interest was the number of items or steps comprising individual tasks on the WTPT. In scoring the WTPT, a person's score on a task is determined by the number of correct steps completed on the task. Items were treated as nested within tasks since they were in fact different for each task on the WTPT. For each task, the items can be considered random samples of larger possible universes of possible items.

Again, the items facet for the WTPT was unbalanced since the number of steps for a task ranged from as little as four to over 30. For each specialty, tasks with as few as three, four, or five items were excluded from the analysis. The next smallest number of items on a task was used as the number of conditions for the items within tasks facet. That number of items was randomly selected from all other tasks included in the design. For example, for the Information Systems Radio Operator, tasks with less than six items were not analyzed. Six items were randomly sampled for all tasks with more than six steps. For two specialties, AFS122x0 and AFS3242x0, after eliminating those tasks with a small number of steps, the remaining tasks were only those which were nested within methods. Consequently, analyses were conducted only on the nested design for these two specialties.

Facets for Substitutability Design

The final generalizability design was used to assess the degree to which the assessment of individuals' proficiency levels were generalizable over the three primary measurement methods: Ratings, WTPT, and job knowledge tests scores. Two analyses were conducted. The first was conducted on AFSs 426x2, 272x0, 328x0, and 492x1. For these, the method facet consisted of two conditions, task-level ratings and overall WTPT scores. In the second analysis, conducted for only the AFSs 122x0, 732x0, 324x0, and 423x5, the method analysis consisted of all three evaluation methods. (Job knowledge scores were also available for these jobs). For all specialties, separate analyses were conducted with ratings by all three rating sources. The results were similar across rating sources, but scores were most generalizable using supervisor ratings. Thus, only these results are presented below.

The second facet was the number of tasks. The number of conditions for the task facet was equal to the smaller number of tasks which constituted either the hands-on or interview component of the WTPT for a specialty (usually about 11). An equivalent number of tasks were randomly sampled from the other WTPT component, from the task-level rating form, and from the job knowledge test. For the job knowledge data, task scores were computed by determining the percentage of questions correct within each task sampled. Tasks were considered either crossed with, or nested within methods, depending on whether the focus was on unique or common tasks.

D Study Analyses

D study analyses were conducted using variance components from the G studies to simulate the treatment of error measurement through multiple operationalizations of instruments. The D study results included variance components for individual effects, total universe score variance (variance due to individual differences, often $\underline{\sigma}^2_p$), relative error variance ($\underline{\sigma}^2_\delta$, equal to the sum of all effects which contain p and at least one other index), absolute error variance ($\underline{\sigma}^2$, equal to the sum of all effects in the design except $\underline{\sigma}^2_p$), and their associated generalizability coefficients ($\underline{\varepsilon P}^2$, for relative decisions; and $\underline{\Theta}$, for absolute decisions).

Conditions in the D study were defined by possible uses of the measures (Gillmore, 1983). Specifically, all facets were treated as random, except for analyses of the WTPT. Then, the methods facet was analyzed as both a random and a fixed facet. Random facets imply that the conditions of a facet represent a random sample from an essentially larger set of possible cases, or that the conditions sampled in the study could be replaced with other elements of some larger set of possible observations without affecting the universe score (Shavelson & Webb, 1981). When a random facet is specified, generalization is not limited to the set of D study conditions, but instead extends to the entire range of admissable observations. In contrast, a fixed facet implies that the conditions observed in the G study exhaust the range of possible conditions of interest to the organization. It also implies that the organization intends to use an average or total score over conditions of the facet.

Secondly, the number of conditions observed for each facet were systematically varied at the D study level to estimate generalizability under measurement conditions of various levels of complexity. For example, G coefficients were computed for the multiple combinations of possible sizes of the WTPT (e.g., 10 items on 10 tasks with one WTPT method, or 15 items on 5 tasks with two methods). Operationally, a D study variance component is adjusted by dividing the variance component by the number of conditions of any facet indicated by its subscript. For example, the G study estimate for $\underline{\sigma}^2_{i:f}$ would be divided by 30 if 10 items on each of three forms was specified as a set of D study conditions.

To distinguish D study estimates from unitary G study values, D study facets were noted by capital letters in the subscript. However, the "p" associated with individuals remains lower-case since persons are not treated as a facet in these analyses. Thus, the G study effect $\underline{\sigma}^2_{i:f}$ is indicated as $\underline{\sigma}^2_{I:F}$ at the D study level, while $\underline{\sigma}^2_{ps}$ is indicated as $\underline{\sigma}^2_{pS}$ (Brennan, 1983; Brennan & Kane, 1979).

III. RESULTS

Ratings Design

Descriptive Results. Tables 1 thru 8 display, within combinations of rating form and rating source, the average item mean and the average scale intercorrelation. Also presented in the tables are averaged correlations indicating convergent validity across sources. These show the correlation between two sources averaged over all items on a form.

Several trends are evident from inspection of these tables. Mean self ratings tend to be slightly higher than mean ratings from peers and supervisors. For example, for Personnel Specialists, mean self ratings ranged from 4.05 to 4.28 across forms, while supervisor ratings ranged from 3.67 to 3.90 and peer ratings ranged from 3.70 to 3.95. This pattern is consistent across all eight specialties, and is similarly observed in nonmilitary contexts as well (Kraiger, 1985).

A second trend is that the average dimension intercorrelation within a form are smaller for self ratings than for those of supervisors or peers. For example, for the Aircrew Life Support speciality, the average for self ratings ranged across forms from .31 to .35, but from .49 to .62 for supervisors and from .49 to .63 for peers. Since the average dimension intercorrelation can be interpreted as an index of halo (Saal, Downey, & Lahey, 1980), the present results suggest that incumbents commit less halo than other sources, or show a greater awareness of their strengths and weaknesses than do supervisors or peers.

Finally, it can be seen that convergent validity coefficients are greater between peers and supervisors than between incumbents and either other source. For example, among Avionic Communications Specialists, the average correlation across dimensions of the Air Force wide forms was .24 between incumbents and either peers or supervisors, but was .38 between peers and supervisors.

While these analyses are useful for gauging certain main effects due to sources, they do not address multivariate effects of measurement facets on ratings. They also do not permit estimation of the relative contributions to error by each facet. Such issues are best addressed in the generalizability analyses presented immediately below.

Table 1. Descriptive Statistics for Rating Variables, for Jet Engine Mechanic

Source:				$\underline{r}^a$ with	
Form	$\underline{m}^a$	$\underline{r}^b$	Self	Supe.	Peer
Self:					
Task	4.02	.30	--	.11	.15
Dimensional	3.80	.41	--	.31	.34
Global	4.13	.38	--	.28	.22
Air Force	3.74	.37	--	.27	.25
Supervisor:					
Task	3.84	.53	.11	--	.13
Dimensional	3.55	.58	.31	--	.40
Global	3.86	.53	.28	--	.51
Air Force	3.51	.58	.27	--	.36
Peer:					
Task	3.94	.49	.15	.13	--
Dimensional	3.66	.55	.34	.40	--
Global	3.80	.41	.22	.51	--
Air Force	3.45	.50	.25	.36	--

[a] averaged across like dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 2. Descriptive Statistics for Rating Variables, for Avionic Communications Specialist

Source:			r[a]	with	
Form	m[a]	r[b]	Self	Supe.	Peer
Self:					
Task	3.99	.60	--	.18	.25
Dimensional	4.03	.40	--	.37	.22
Global	4.04	.09	--	.31	.18
Air Force	3.79	.63	--	.24	.24
Supervisor:					
Task	3.95	.51	.18	--	.26
Dimensional	3.89	.49	.37	--	.40
Global	3.83	.21	.31	--	.38
Air Force	3.63	.43	.24	--	.38
Peer:					
Task	3.87	.42	.25	.26	--
Dimensional	3.95	.61	.22	.40	--
Global	3.86	.45	.18	.38	--
Air Force	3.59	.52	.24	.38	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 3. Descriptive Statistics for Rating Variables, for Air Traffic Control Operator

Source:			r[a]	with	
Form	M[a]	r[b]	Self	Supe.	Peer
Self:					
Task	4.04	.32	--	.22	.32
Dimensional	3.97	.41	--	.24	.25
Global	4.04	.46	--	.18	.21
Air Force	3.89	.39	--	.14	.15
Supervisor:					
Task	3.64	.45	.22	--	.26
Dimensional	3.60	.56	.24	--	.35
Global	3.69	.41	.18	--	.38
Air Force	3.52	.48	.14	--	.24
Peer:					
Task	3.88	.47	.32	.26	--
Dimensional	3.86	.49	.25	.35	--
Global	3.87	.51	.21	.38	--
Air Force	3.68	.43	.15	.24	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 4. Descriptive Statistics for Rating Variables, for Information Systems Radio Operator

Source:			$\underline{r}^a$	with	
Form	$\underline{m}^a$	$\underline{r}^b$	Self	Supe.	Peer
Self:					
Task	4.23	.44	--	.36	.35
Dimensional	4.22	.50	--	.28	.29
Global	4.24	.28	--	.25	.31
Air Force	4.03	.41	--	.24	.14
Supervisor:					
Task	4.29	.49	.36	--	.28
Dimensional	4.16	.51	.28	--	.30
Global	4.06	.37	.25	--	.39
Air Force	3.78	.48	.24	--	.23
Peer:					
Task	4.25	.38	.35	.28	--
Dimensional	4.17	.56	.29	.30	--
Global	4.08	.31	.31	.39	--
Air Force	3.84	.48	.14	.23	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 5. Descriptive Statistics for Rating Variables, for Aircrew Life Support

Source:			$\underline{r}^a$	with	
Form	$\underline{M}^a$	$\underline{r}^b$	Self	Supe.	Peer
Self:					
Task	3.97	.34	--	.25	.25
Dimensional	3.86	.35	--	.23	.22
Global	4.12	.31	--	.23	.26
Air Force	3.84	.33	--	.21	.20
Supervisor:					
Task	3.81	.53	.25	--	.35
Dimensional	3.73	.49	.23	--	.27
Global	3.87	.62	.23	--	.25
Air Force	3.63	.58	.21	--	.15
Peer:					
Task	3.78	.51	.25	.35	--
Dimensional	3.73	.49	.22	.37	--
Global	3.81	.63	.26	.25	--
Air Force	3.57	.53	.20	.15	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 6. Descriptive Statistics for Rating Variables, for Personnel Specialist

Source:			r^a	with	
Form	m^a	r^b	Self	Supe.	Peer
Self:					
Task	4.20	.23	--	.21	.32
Dimensional	4.21	.30	--	.10	.16
Global	4.28	.21	--	.17	.26
Air Force	4.05	.37	--	.13	.17
Supervisor:					
Task	3.82	.31	.21	--	.23
Dimensional	3.77	.44	.10	--	.23
Global	3.90	.52	.17	--	.32
Air Force	3.67	.53	.13	--	.25
Peer:					
Task	3.94	.28	.32	.23	--
Dimensional	3.95	.42	.16	.23	--
Global	3.95	.30	.26	.32	--
Air Force	3.70	.44	.17	.25	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 7. Descriptive Statistics for Rating Variables, for Equipment Laboratory Specialist

Source:			r^a	with	
Form	m^a	r^b	Self	Supe.	Peer
Self:					
Task	3.70	.33	--	.24	.29
Dimensional	3.83	.34	--	.18	.21
Global	3.79	.29	--	.29	.26
Air Force	3.68	.30	--	.31	.24
Supervisor:					
Task	3.49	.50	.24	--	.28
Dimensional	3.61	.49	.18	--	.25
Global	3.60	.45	.29		.29
Air Force	3.49	.48	.31	--	.35
Peer:					
Task	3.59	.41	.29	.28	--
Dimensional	3.72	.55	.21	.25	--
Global	3.72	.37	.26	.29	--
Air Force	3.63	.39	.24	.36	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

Table 8. Descriptive Statistics for Rating Variables, for Aerospace Ground Equipment

Source:			$\underline{r}^a$	with	
Form	$\underline{m}^a$	$\underline{r}^b$	Self	Supe.	Peer
Self:					
Task	3.63	.32	--	.20	.23
Dimensional	3.54	.36	--	.25	.28
Global	3.81	.45	--	.31	.24
Air Force	3.63	.30	--	.31	.26
Supervisor:					
Task	3.38	.51	.20	--	.24
Dimensional	3.30	.59	.25	--	.30
Global	3.51	.63	.31	--	.36
Air Force	3.35	.56	.31	--	.32
Peer:					
Task	3.49	.46	.23	.24	--
Dimensional	3.48	.53	.28	.30	--
Global	3.61	.56	.24	.36	--
Air Force	3.42	.47	.26	.32	--

[a] averaged across dimensions within form.

[b] average dimension intercorrelations within forms and sources.

G Study Results. Summary G study results for analyses of the rating data are presented in Table 9. Variance components for each effect are presented for all eight specialties. Complete G study results for the first four specialties are presented in Appendix A of Kraiger (1990), while complete results for the latter four are presented in Appendix A of this document. The tables in the appendices show the estimated variance components along with their associated degrees of freedom, mean squares, and confidence intervals. The confidence intervals indicate the precision in estimation of the population values of variance components, given the sample size and design complexity. The confidence intervals are based on the ratio of the estimated variance component to its standard error and were calculated from procedures detailed by Satterthwaite (1941, 1946).

The estimated G study variance components in Table 9 indicate that results were similar over occupational specialties. Relatively large variance components are undesirable for all effects but σ^2_p, variability due to individual differences. In all specialties, the largest variance component was the residual term ($\sigma^2_{ps(i:f)}$), ranging from .285 for Air Traffic Control Operators to .395 for Personnel Specialists. Likewise, the σ^2_{ps} term was the second largest estimate in each design, ranging between .140 to .208. The σ^2_p term, universe score variance, is the third largest term for all specialties except Information Systems Radio Operators and Personnel Specialists, and ranged from .047 to .151. Similar narrow ranges across specialties can be seen for the other terms. Only a few terms show considerable variation across specialties. The main effect for rater sources, σ^2_s is near zero in six specialties, but substantially larger for Air Traffic Control Operators and Personnel Specialists. As can be deduced from Tables 3 and 6, this effect the latter specialties was largely due to low mean supervisory ratings for Air Traffic Control Operators and exceptionally high self ratings by the Personnel

Table 9. Estimated Variance Components for G Study of Rating Variables with Three Forms

	Job:							
	JEM	ACS	ATC	ISRO	ALS	PS	PMEL	AGE
Effect	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$
p	.151	.120	.118	.133	.088	.047	.087	.122
s	.015	.015	.036	.001	.010	.041	.010	.016
f	.001	-.001	-.017	-.009	.001	.002	-.005	-.006
i:f	.015	.031	.040	.025	.039	.045	.049	.054
ps	.186	.173	.208	.173	.193	.172	.140	.160
pf	-.003	-.030	-.009	.023	.028	.023	.027	.022
sf	.001	-.008	.000	.003	.000	.000	-.001	.000
psf	.016	-.018	.010	.036	.061	.043	.033	.048
p(i:f)	.057	.106	.066	.089	.074	.094	.065	.055
s(i:f)	.004	.019	.000	.002	.005	.005	.002	.007
ps(i:f)	.293	.330	.285	.306	.353	.395	.322	.359

Specialists. Also, the $\underline{\sigma}^2_{psf}$ term, indicating the extent to which ratees were differentially ranked by sources depending on which form was used, was considerably lower for three specialties, 426x2, 272x0, and 328x0, than in the other five. This pattern suggests that in these three specialties, ratees were ranked similarly regardless of which combination of rater source and form was used, but in the other five specialties, the interaction of form and source affected a ratee's relative ranking. For example, an incumbent in Aerospace Ground Equipment might be ranked above a co-worker by peers using one form, but ranked below by a supervisor using a different form.

D Study Results. D study analyses of the rating data were based on analyses of the three-form analyses. Complete results of these analyses are presented for the first four specialties in the appendix of Kraiger (1990) and for the latter four specialties in Tables A-5 to A-8 of Appendix A of this document. In addition, summary G coefficients for relative decisions ($\underline{\varepsilon P}^{2}$) for all specialties are presented in Figure 1 for two sets of measurement conditions: A single source using a single 8-item form (representing typical organizational operationalizations of rating methods), and three sources using four 8-item forms (the D study which best approximates the actual measurement conditions on the JPM). The generalizability coefficient represents the proportion of observed score variance which is attributable to universe score variance or individual differences. An examination of estimated variance components within specialties provides evidence of desirable or undesirable measurement effects under particular rating conditions, while an examination of the summary G coefficients indicate the overall dependability of measures under those conditions.

As shown in Figure 1, rating measures are more reliable when ratings are averaged over multiple sources and multiple forms. With a single source using a single 8-item form, G coefficients ranged between .135 and .302. In contrast, by averaging scores over all three sources and four forms, the generalizability scores ranged from .388 to .641, with most values above .500. While these latter values are still below recommended values by Cardinet, Tourneur, and Allal (1976), they may be acceptable for some uses of the rating data. Notably, the G coefficients are comparable across the specialties, except that the values for Personnel Specialists were considerably lower than those in the other seven specialties.

Inspection of the full D study analyses in the appendices yields insights into the increases in generalizability with increased numbers of rating dimensions, forms, and (particularly) sources. For example, the $\underline{\sigma}^{2}_{p(i:f)}$ term is small, but non-trivial in the G study results presented above. By averaging ratee scores over multiple items and/or forms, this undesirable source of variance can be virtually eliminated at the D study level. Similarly,

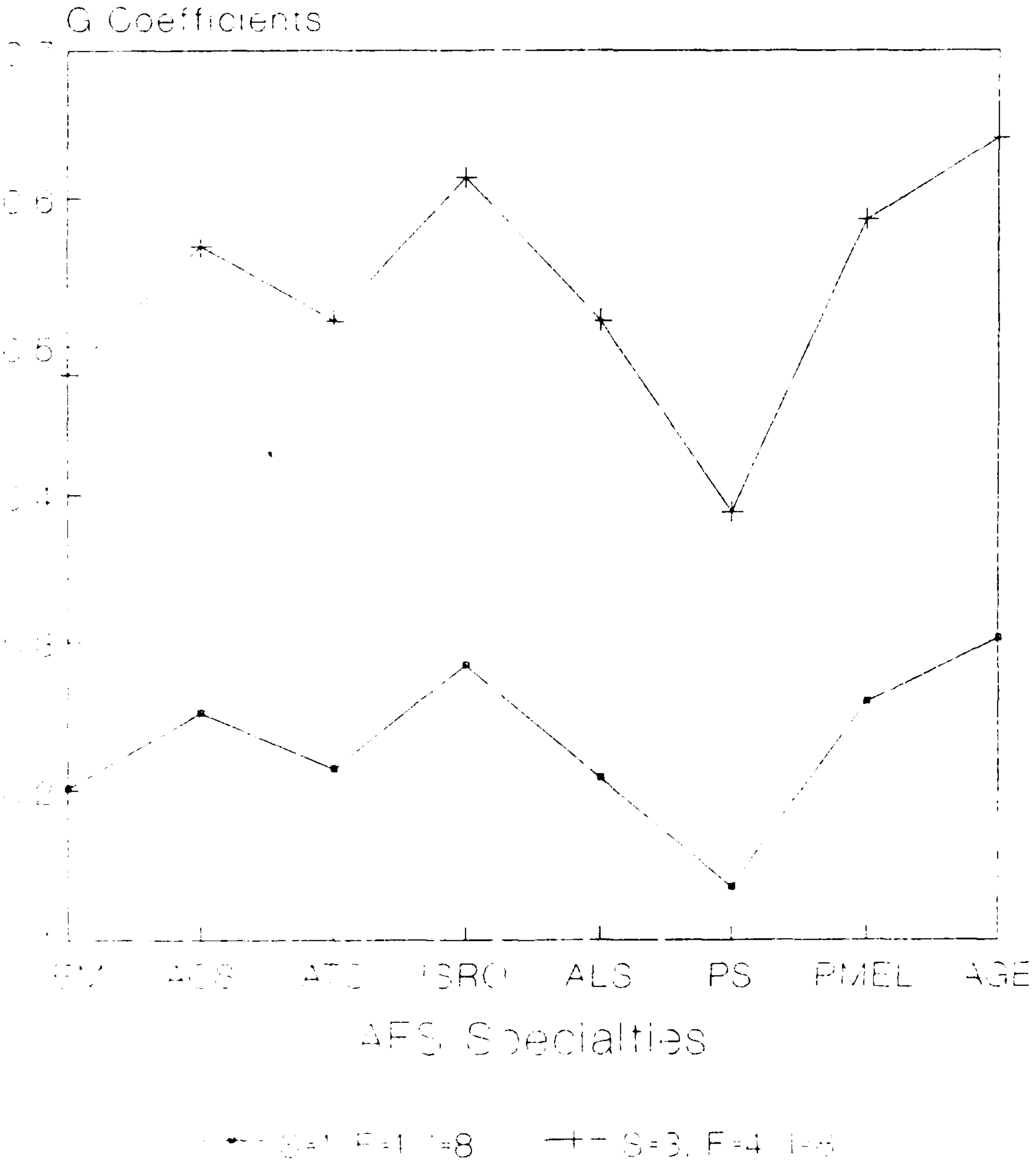

Figure 1. G Coefficients for Performance Rating Data for Eight Occupational Specialties

averaging over multiple sources reduces the $\underline{\sigma}^2_{ps}$ and $\underline{\sigma}^2_{psf}$ terms, though the ratee-by-source interaction still contributes considerable variance to the rating design, even when ratings are averaged over three sources. This source of variance is the greatest threat to the generalizability of the performance ratings. Finally, it should be noted that individual estimated D study variance components were quite similar over occupational specialties.

Within Source Analyses

Because of the large effect for the interaction of persons and sources, a set of secondary analyses were performed within each rater source for each specialty. In these analyses, facets of interest were forms and items within forms. All analyses employed the three-form design. Both G and D study results for these analyses are displayed in Table 10. A D study generalizability coefficient is presented only for a single condition -- ratings on a single 8-item form. This generalizability coefficient is also displayed in Figure 2 for each source.

Again, the results were marked by consistency across specialties, for both estimated variance components and G coefficients. The largest source of variance was typically the interaction of persons and items within forms ($\underline{\sigma}^2_{p(i:f)}$), a term confounded within random error ($\underline{\sigma}^2_{e}$). Variance due to individual differences, $\underline{\sigma}^2_{p}$ was also substantial for each source within each specialty, while all other sources of variance were negligible.

In contrast to the prior results, fairly large D study generalizability coefficients were obtained, even under less rigorous measurement specifications (i.e., a single 8-item form). The majority of G coefficients under these conditions ranged from .660 to .750 across sources and specialties. Within the Jet Engine Mechanic and Information Systems Radio Operator specialties, the largest generalizability coefficient was found for the supervisory ratings ($\underline{E}\underline{\rho}^2$ = .728, .726 respectively), while for Avionic Communications

Table 10. G and D Study Results for Within Source Analyses

	Job:							
Source:	JEM	ACS	ATC	ISRO	ALS	PS	PMEL	ACT
Effect	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$	$\underline{\sigma}^2$
Self:								
p	.192	.161	.218	.219	.145	.149	.147	.163
f	.035	.014	.011	.030	-.006	.001	-.007	-.008
i:f	.025	.034	.021	.019	.066	.034	.062	.087
pf	.048	.030	.038	.073	.094	.034	.044	.065
p(i:f)	.351	.415	.376	.314	.473	.451	.403	.464

$\underline{\varepsilon P}^2$ when								
$\underline{f}$=1, $\underline{i:f}$=8	.666	.665	.720	.660	.496	.622	.609	.572
Supervisor:								
p	.375	.275	.312	.289	.363	.271	.279	.400
f	.026	.038	.014	.062	.002	-.006	-.007	-.005
i:f	.026	.026	.029	.035	.031	.068	.050	.049
pf	.097	.069	.103	.063	.087	.070	.090	.061
p(i:f)	.346	.420	.400	.373	.396	.456	.382	.383

$\underline{\varepsilon P}^2$ when								
$\underline{f}$=1, $\underline{i:f}$=8	.728	.694	.671	.726	.727	.691	.670	.796
Peer:								
p	.265	.357	.291	.282	.337	.234	.256	.282
f	.047	.051	.019	.056	.006	.010	-.004	-.005
i:f	.024	.017	.031	.015	.035	.047	.042	.046
pf	.077	.020	.075	.072	.088	.093	.047	.083
p(i:f)	.350	.328	.387	.314	.411	.562	.374	.396

$\underline{\varepsilon P}^2$ when								
$\underline{f}$=1, $\underline{i:f}$=8	.687	.853	.703	.716	.707	.599	.732	.690

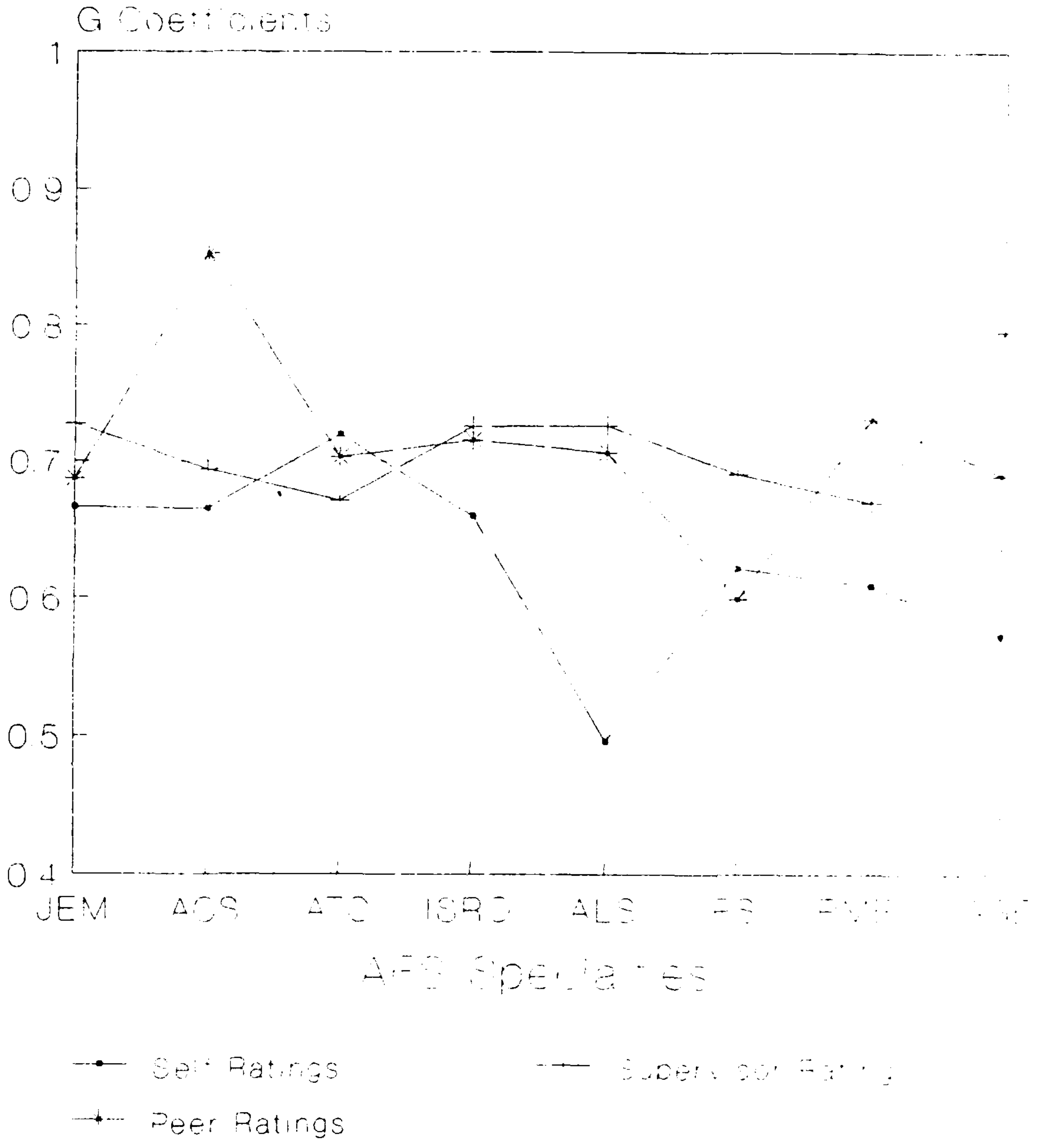

Figure 2. εP^2 Within Rating Sources for Eight Occupational Specialties

Specialists the largest coefficient was found for peer ratings (εP^2=.853), and for Air Traffic Control Operators the largest coefficient was found for self ratings (εP^2=.720).

G Study Results, WTPT Data

Results of the G study analyses across specialties are presented in Tables 11 (for the crossed design) and 12 (for the nested design). Tables A-13 through A-20 in the appendix display mean squares, variance components, and confidence intervals for each effect in both designs, shown separately by specialty.

Results for the crossed design (Table 11) indicate considerably greater variability across specialties than was seen with the rating data. For example, variance due to individual differences, σ^2_p, ranged from .006 for Avionics Communications Specialists to .032 for Information Systems Radio Operators. Likewise, the residual term, $\sigma^2_{pm(i:t)}$, was considerably larger in the Jet Engine Mechanic than in the other three specialties. The σ^2_{pm} and σ^2_{pmt} terms were relatively small and consistent across specialties, but considerable variation in estimates was found for the σ^2_{pt} and $\sigma^2_{p(i:t)}$ terms. The estimate for the person by task interaction was near zero for Jet Engine Mechanics, but substantially larger in the other three specialties. This indicates that incumbents in these latter three specialties were differentially ranked on performance, depending on the task. The greatest variability was found for the interactions of persons and items nested with tasks. This term was again near zero for Jet Engine Mechanics, substantially larger for Avionics Communication Specialists and Information Systems Radio Operators, and larger yet for Air Traffic Control Operators. In absolute terms, the estimated variance component $\sigma^2_{p(i:t)}$ for Avionic Communications Specialties and Information Systems Radio Operators was about five times greater than the estimate for

Table 11. Estimated Variance Components for G Study of WTPT Variables for Tasks Crossed with Methods[a]

	Job:					
	JEM	ACS	ATC	ISRO	PS	AGE
Effect	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2
p	.008	.006	.007	.032	.019	.006
m	.013	.014	.000	.000	.003	-.001
t	.000	.016	.008	.007	-.005	.004
i:t	.000	.017	.010	.005	-.003	-.008
mt	.001	.000	.000	.000	.016	.022
pm	.002	.007	.001	.000	-.014	.000
pt	.008	.025	.034	.028	-.014	.001
p(i:t)	.009	.032	.073	.012	.000	-.004
pmt	.012	.008	.007	.020	.078	.020
m(i:t)	.029	.014	.009	.002	.021	.063
pm(i:t)	.127	.074	.065	.052	.094	.149

[a] For the jobs of Aircrew Life Support and Precision Measurement Equipment Laboratory Specialist, the only tasks remaining after eliminating tasks with a small number of steps were those nested within methods. Consequently, the analyses in this table were conducted only on the nested design for these two specialties.

Table 12. Estimated Variance Components for G Study of WTPT Data for Tasks Nested in Methods

	Job:							
	JEM	ACS	ATC	ISRO	ALS	PS	PMEL	AGE
Effect	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2
p	.008	.013	.007	.029	.018	.038	.004	.011
m	.013	.001	-.001	-.001	.004	-.007	.004	-.006
t:m	.003	.014	.012	.008	.026	.013	.010	.036
i:t:m	.020	.030	.032	.009	.037	.008	.037	.053
pm	.001	-.001	-.002	-.003	-.001	-.031	-.001	-.006
p(t:m)	.019	.032	.018	.051	.027	.051	.011	.037
p(i:t:m)	.144	.108	.128	.030	.119	.078	.095	.126

Air Traffic Control Operators and 15 times greater than the corresponding estimate for Jet Engine Mechanics.

Results for the design with tasks nested in methods (Table 12) were similar to those of the crossed design. There was considerable variation across jobs in $\sigma^2_{t:m}$ and $\sigma^2_{i:t:m}$, but little variation in σ^2_{pm}.

These low variance components for the person-by-method interaction indicated that incumbents were not differentially ranked by their performance on the two WTPT methods (hands-on and interview). The residual term, $\sigma^2_{p(i:t:m)}$ was the largest variance component for each specialty, though the values of this term varied over specialty. Finally, there was also considerable variation in the $\sigma^2_{p(t:m)}$ term, with estimates being substantially lower in the Jet Engine Mechanic and Air Traffic Control Operator specialties than in the other two. Thus, only in these two specialties were incumbents not differentially ranked by particular tasks.

D Study Results, WTPT Data

D study analyses were based on the crossed design, since this design permitted assessment of a greater number of effects. D study results for each specialty are displayed graphically in Figure 3 and in tabular form in Tables A-15 through A-16 of Appendix A.

Unlike the D study results for the rating data, changes in specifications of measurement conditions produced considerable variations in the resulting generalizability curves. Using both the hands-on and interview components reduces the associated variance components and improves the generalizability of WTPT scores. In general, scores averaged over both methods using a small number of items and a small number of tasks were more generalizable than scores on a single method with a substantially greater number of tasks or items.

Inspection of Figure 3 reveals that the greatest levels of generalizability were obtained for Information Systems Radio Operators, Personnel Specialists, and Aerospace Ground Equipment incumbents. For these specialties, G coefficients above .750 can be obtained with 15 tasks, each with 10 steps, assessed by both hands-on and interview formats. G coefficients were considerably lower in the other specialties. The lowest levels of generalizability occurred for Avionic Communications Specialists. Even with scores averaged over two methods, 15 tasks, and 10 steps, $\underline{\varepsilon P}^2$ equaled only .504. Generalizability levels were somewhat higher for the Air Traffic Control Operators. $\underline{\varepsilon P}^2$ equalled .683 under similar measurement conditions. It is clear that for these specialties, the WTPT should be constructed with as many items and tasks as feasible. It is also worth noting that generalizability coefficients varied over occupations, making overall conclusions about the dependability of the WTPT more tenuous.

G and D Study Results, Substitutability Design

G study estimated variance components, as well as D study estimates of $\underline{\varepsilon P}^2$ for the substitutability design are

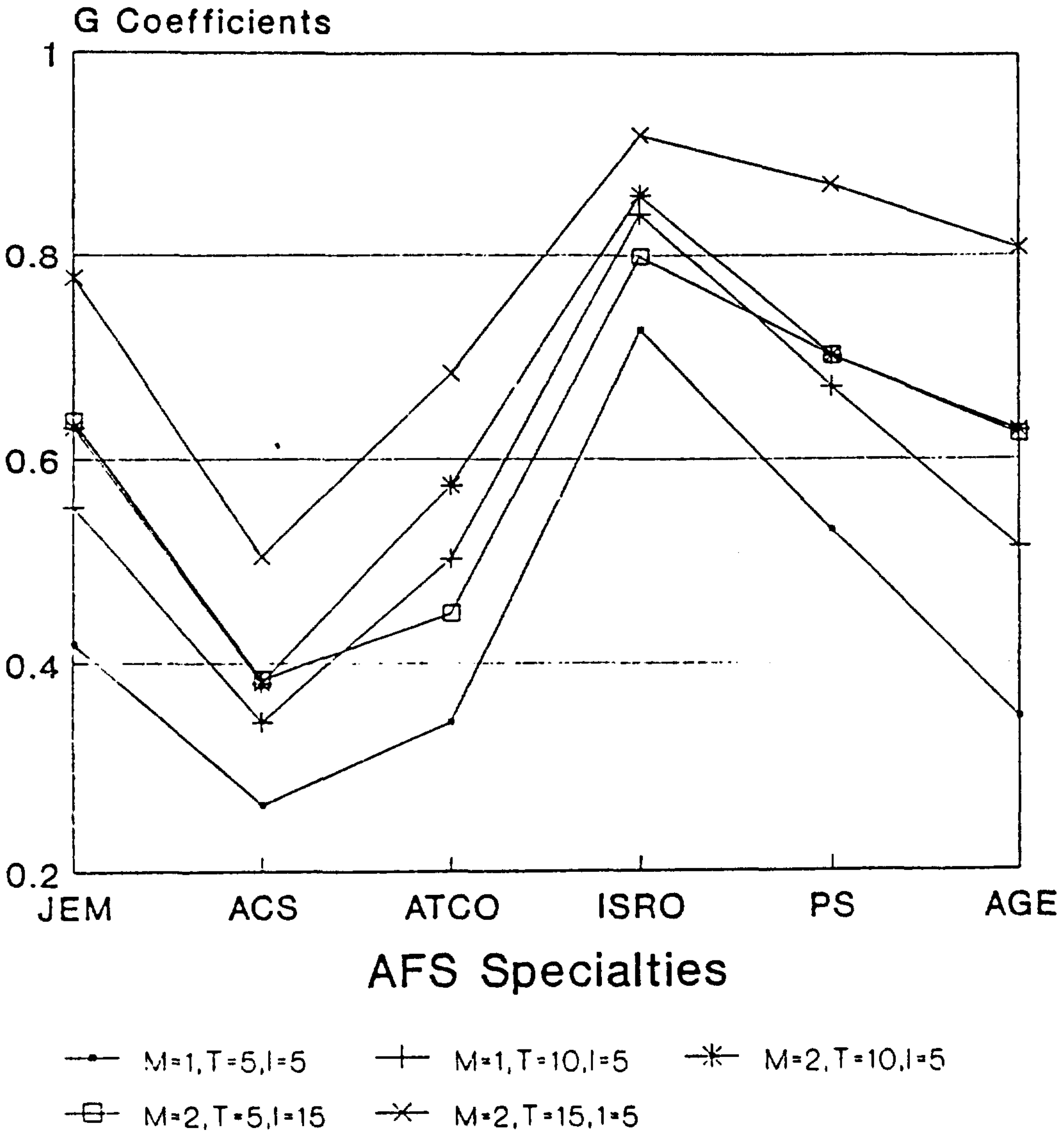

Figure 3. G Coefficients for WTPT Scores for Six Occupational Specialties

presented in Table 13. The number of methods assessed at the G study level vary by specialty. For the first four specialties (in the Table), the generalizability of scores was assessed across proficiency ratings and WTPT scores; for the latter three specialties, generalizability was assessed across ratings, WTPT scores, and job knowledge test scores for the latter three columns. For each analysis, supervisory ratings were used for the rating data. The lower portion of Table 13 also presents D study results for two sets of measurement conditions: A single method of assessing 15 tasks and scores averaged over all three methods, each assessing 15 tasks.

In no instance are performance scores generalizable over the evaluation methods. The greatest level of generalizability was obtained for the Information Systems Radio Operator and Aircrew Life Support specialties ($\underline{\varepsilon P}^2$ = .491 and .439, respectively), but even these values are well below acceptable levels. In general, only a little over a third of the observed variance in individuals' scores can be attributed to universe score variance (or individual differences). Looking at the individual variance components, it is clear that the low G coefficients are the result of large values for the $\underline{\sigma}^2{}_{pm}$ and residual terms. The large values for $\underline{\sigma}^2{}_{pm}$, which can be reduced by a third, at the most, at the D study level, indicate that incumbents are differentially ordered by methods, a strong threat to the generalizability of the system.

(The high estimates for $\underline{\sigma}^2{}_{m}$ indicate large mean differences between methods. This is an artifact produced by a 5-point scale used for the ratings, a 1-point scale used for the two WTPT methods and a .00 to 1.00 scale used for the job knowledge tests.)

IV. DISCUSSION

The purpose of the present investigation was to apply G theory to the data collected on the Air Force Performance Measurement Project in order to address the following issues:

Table 13. G and D Study Results for Substitutability Design using Supervisor Ratings, WTPT Scores, and Job Knowledge Test Scores

	Job:						
	JEM	ACS	ATC	ISRO	ALS	PS	AGE
Effect	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2
Persons (p)	.016	.012	.007	.031	.617	.087	.063
Methods (m)	3.202	3.196	2.801	4.219	7.767	12.054	6.374
Tasks (t)	-.002	.002	.006	.002	.068	.089	.150
mt	.033	.021	.042	.023	.344	.404	.974
pm	.130	.126	.149	.086	2.245	.327	.323
pt	-.002	.002	.006	.002	.023	.085	.022
pmt	.144	.188	.181	.137	1.900	2.697	1.882
εP^2 when:							
m=1, t=15	.104	.076	.044	.244	.207	.178	.126
m=3, t=15	.259	.198	.120	.491	.439	.393	.301

Note: For the three right-hand columns, tasks were treated as nested within methods, so that the row values are t:m, not t; and p(t:m) not pmt.

The psychometric adequacy of the ratings, the psychometric adequacy of the WTPT, and the degree to which the ratings and/or job knowledge test scores are acceptable surrogates for the WTPT. Also of interest are whether data relevant to the above questions are consistent across specialties, and whether particular measurement technologies can be reduced in scope without compromising the dependability of scores. Each of the issues are addressed below, along with recommendations regarding the JPM project.

Psychometric Quality of Performance Ratings

Evidence for the psychometric quality of the performance ratings comes from G and D study results within each occupational specialty. Cardinet et al. (1976) recommended

.80 as a minimally acceptable level for G coefficients. Given this value, the generalizability levels of proficiency ratings for relative decisions are inadequate in each specialty, regardless of the measurement conditions specified. However, the benchmarks of Cardinet et al. were offered principally for paper-and-pencil tests, and it is logical to expect G coefficients for rating systems to be lower. The suitability of any generalizability coefficient should be interpreted within the context of results from similar studies.

Given these qualifications, it is reasonable to be somewhat optimistic about the fidelity of the proficiency ratings. For six of the specialties, G coefficients are greater than .70 when scores are averaged over three sources, at least two forms, and at least eight items. Generalizability coefficients for the other two specialties are only slightly lower. This indicates that under such measurement conditions, about half the observed variance in scores can be attributed to individual differences. These G coefficients are about the same as, or greater than, coefficients reported in similar rating studies by McHenry, Hoffman, and White (1987) and Webb and Shavelson (1987). Further, they are higher than typical inter rater reliability estimates (King, Hunter, & Schmidt, 1980).

The relatively high variance components within sources, coupled with the large $\underline{\sigma}^2_{ps}$ term, suggest that ratings are very dependable within source, but differ considerably in how ratees are ranked across sources. Other researchers have questioned whether ratings from different sources should be expected to converge, since different sources may have different opportunities to observe ratee performance, or vary in their interpretation of behavior (Borman, 1974; Guion, 1966; Klimoski, & London, 1974). More recently, Kraiger and Teachout (1990) have openly questioned the expectation of agreement over rating sources, and have called for meaningful research on the nature of these differences. Operationally, the implication of these results are that the Air Force should continue collecting and averaging scores over sources to reduce error variance at the D study level. It should also initiate research to understand why sources do vary in their assessment, and whether ratings of one source are more valid than others.

Finally, it is noted that results are very consistent across the eight specialties studied. There appears to be little or no variability across jobs in the psychometric characteristics of the rating system. Thus, there is less of a need to continue collecting and assessing rating data in additional specialties, unless attempts are made (and tested) to increase convergence across sources.

Psychometric Quality of WTPT Scores

Evidence of the psychometric quality of the WTPT method comes from G and D study results within each occupational specialty. In contrast to the rating data, there is considerably greater variability across specialties for the WTPT data. Acceptable levels of generalizability are reached under a variety of measurement conditions for all jobs except Avionic Communications Specialties. For this AFS, D study generalizability coefficients are well below .50 under all measurement conditions studied.

Inspection of the G study estimates of variance components reveals that the interaction of persons and tasks ($\underline{\sigma}^2_{pt}$ or $\underline{\sigma}^2_{p(t:m)}$) the interaction of persons and items ($\underline{\sigma}^2_{p(i:t)}$ or $\underline{\sigma}^2_{p(i:t:m)}$) and the residual term ($\underline{\sigma}^2_{pm(i:t)}$ or $\underline{\sigma}^2_{p(i:t:m)}$ all contributed substantial error variance in different combinations of jobs or designs. The interaction of persons and tasks contributed a substantial portion of variance in most specialties. This indicates that persons were differentially ranked in terms of performance on tasks. There are a number of potential reasons for this, including base-to-base differences in mission, airman specialization, or on-the-job training. For example, suppose Airman A is assigned to perform task 1 but not task 2, while Airman B is assigned to perform task 2 but not task 1. These two Airman will be differentially ranked on these two tasks, even if there are no true differences in job proficiency.

The residual terms are somewhat high, especially for Jet Engine Mechanics and for Air Traffic Control Operators when tasks are treated as nested within methods. For the latter job, this residual term includes the confounding of the $\underline{\sigma}^2_{p(i:t:m)}$ and $\underline{\sigma}^2_{e}$ terms. Since the interaction of persons and items was large for this specialty in the crossed design, it is safe to assume that the $\underline{\sigma}^2_{p(i:t:m)}$ term accounts for much of the variance in the residual term for the reasons speculated above. For Jet Engine Mechanics though, the residual term is large in both designs. For the crossed design, the residual term confounds $\underline{\sigma}^2_{pm(i:t)}$ and $\underline{\sigma}^2_{e}$. Since other terms containing the interaction of persons and methods in this design are very small ($\underline{\sigma}^2_{pm}$ and $\underline{\sigma}^2_{pmt}$), it can be reasonably assumed that it is the effects of $\underline{\sigma}^2_{e}$ which results in the extremely high residual term for this specialty. Undifferentiated error includes both random error and other systematic effects not included in the design. For example, if persons were differentially ranked by test administrators, or persons from various bases were differentially ranked, these effects would be reflected by the residual term, but could not be assessed by the present design. At best, Air Force decision makers could intuitively judge whether it is plausible to assume that administrators, bases, or other systematic effects were more problematic with the Jet Engine Mechanic specialty than others. On the other hand, since this was the first specialty in which the WTPT was designed and applied to data collection, decision makers may also wish to judge whether it is likely that there was greater random error introduced through the process of developing the procedures.

It should also be noted that the variance component for the persons-by-methods interaction ($\underline{\sigma}^2_{pm}$) was extremely small in both designs, for all specialties. This means that testtakers were ranked the same whether they were actually performing the task or merely describing it. Thus, the interview format appears to be an acceptable substitute for the more expensive hands-on component.

Finally, the variability in variance components and generalizability coefficients across specialties is re-emphasized. Additional data collection in other specialties may be warranted, though the trend of findings to date is positive.

Other Measures as Surrogates for WTPT Scores

Evidence of the adequacy of proficiency ratings and job knowledge test scores as surrogates of the WTPT comes from G and D studies of the substitutability design. Regardless of whether scores are averaged across sources, or considered for each source by itself, there is very little convergence between ratings and WTPT scores, or ratings and WTPT and job knowledge scores. Thus, task proficiency ratings are not adequate substitutes for the WTPT.

One question which follows is which set of scores is the more trustworthy. Under normal measurement conditions, the generalizability analyses discussed above indicate that the performance ratings are more dependable for Avionic Communications Specialists, Aerospace Ground Equipment Operators, and Air Traffic Control Operators, but that WTPT scores are more dependable for Jet Engine Mechanics, Personnel Specialists, and Information Systems Radio Operators. Such conclusions are tempered by the confidence one has that all measurement conditions which might affect scores were included in analyses of the ratings and WTPT scores. For example, if test administrators did contribute significant error variance to WTPT scores, designs which permitted estimation of such effects could have resulted in superior G coefficients for WTPT tests in all specialties. These sources of variability can not be directly estimated because the factors in question were not allowed to vary in the specialties studied to date. Future research efforts may attempt to assess these other factors. At present though, there appears to be no reason to favor one methodology over the others and the wisest course of action would seem to be to continue using all sets of scores in decisions.

Recommendations

(1) There appears to be little utility in collecting additional information on proficiency ratings for purposes of understanding their psychometric quality. Results to date are very consistent across the specialties already studied. The best reason to continue studying ratings data would be to test differences in aspects of scale development or data collection (e.g, variations in rater training programs). From a research perspective, it would be valuable to continue exploring the differential meaning and validity of ratings by different sources.

(2) Proficiency ratings appear to be adequate criteria for validation purposes and the methodology developed in these specialties should be applied in others as well. It is possible to reduce the number of forms to one or two and maintain current fidelity levels, but ratings should be collected from (and averaged over) all three sources.

(3) The WTPTs should be applied in other specialties for both pure research and validation purposes. Additional research is needed because the expected generalizability coefficients or relative size of individual variance components cannot be extrapolated from the data collected to date. In general though, the data analyses presented above suggest that the WTPT is the single best method of evaluating incumbent performance for the purpose of validating the Armed Services Vocational Aptitude Battery.

(4) It is unwise to consider proficiency ratings or job knowledge test scores as substitutes for the WTPT. Instead, they each appear to represent vastly different aspects of the total criterion space. There is little overlap in the substantive universes assessed by each. Thus, all three measures can be considered "correct," even though they are

essentially unrelated. Other research strategies which emphasize comparing both sets of scores to other indicators or predictors of performance appear to be necessary to understand the latent constructs measured by each (Borman, 1987).

REFERENCES

Bentley, B.A., Ringenbach, K.L., & Augustin, J.W. (1989, May). *Development of Army job knowledge tests for three Air Force specialties* (AFHRL-TP-88-11, AD-A208 245). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Borman, W.C. (1974). The ratings of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance*, *12*, 105-124.

Borman, W.C. (1987, April). Comments as panelist in M.M. Kavanagh (Chair), *A roundtable discussion of research issues in criterion measurement*. Annual meeting of the Society for Industrial/Organizational Psychology, Atlanta, GA.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brennan, R.L., & Kane, M.T. (1979). Generalizability theory: A review. In L.J. Fryans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications*. San Francisco: Jossey-Bass.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, *13*, 119-135.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Gillmore, G.M. (1983). Generalizability theory: Application to program evaluation. In L.J. Fryans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications. San Francisco: Jossey-Bass.*

Guion, R.M. (1966). *Personnel testing*. New York: McGraw-Hill.

Hedge, J.W., & Teachout, M.S. (1986, November). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37, AD-A174 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

King, L.M., Hunter, J.E., & Schmidt, F.L. (1980). Halo in a multidimensional forced choice performance evaluation scale. Journal of Applied Psychology, 65, 507-516.

Klimoski, R.J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59, 445-451.

Kraiger, K. (1985, September). Analysis of relationships among self, supervisory, and peer ratings of performance. Final report submitted to the AFOSR and AFHRL, Brooks AFB, TX.

Kraiger, K. (1989, April). Generalizability theory: An assessment of its relevance to the Air Force job performance measurement project (AFHRL-TP-87-70, AD-A207-107). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kraiger, K. (1990, April). Generalizability of performance measures across four Air Force specialties (AFHRL-TP-89-60). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kraiger, K., & Teachout, M.S. (1987, April). Generalizability theory as evidence of the construct validity of ratings. In G. Laabs (Chair), Applications of generalizability theory to military performance measurement. Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC.

Kraiger, K., & Teachout, M.S. (1990). Generalizability theory as construct-related evidence of construct validity of job performance ratings. Human Performance, 3, 19-35.

McHenry, J.J., Hoffman, R.G., & White, L.A. (1987, April). A generalizability analysis of peer and supervisory ratings.

In G. Laabs (Chair), Applications of generalizability theory to military performance measurement. Symposium at the annual meeting of the American Educational Research Association, Washington, DC.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309-316.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.

Searle, S.R. (1971). *Linear models. New York: Wiley.*

Shavelson, R.J. (July, 1986). *Generalizability of military performance measurements: I. Individual performance*. Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy of Sciences, Washington, DC.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, *31*, 133-161.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932.

Webb, N., & Shavelson, R. (1987, April). Generalizability theory and job performance measurement. In G. Laabs (Chair), *Applications of generalizability theory to military performance measurement*. Symposium at the annual meeting of the American Educational Research Association, Washington, DC.

APPENDIX A: ADDITIONAL G AND D STUDY RESULTS WITHIN OCCUPATIONAL SPECIALTIES

Table A-1. Estimatec Variance Components for Aircrew Life Support, Three-form Analysis

Effect	Df	Ms	σ^2	90% Confidence Intervals
Persons (p)	216	9.70	.088	$.058<\sigma^2<.118$
Sources (s)	2	43.15	.010	$.000<\sigma^2<.023$
Forms (f)	2	30.27	.001	$.000<\sigma^2<.011$
Items within f (i:f)	15	27.10	.039	$.015<\sigma^2<.063$
ps	432	4.20	.193	$.167<\sigma^2<.219$
pf	432	1.45	.028	$.018<\sigma^2<.038$
sf	4	1.61	.000	$.000<\sigma^2<.000$
psf	864	.72	.061	$.051<\sigma^2<.071$
p(i:f)	3,240	.57	.074	$.065<\sigma^2<.083$
s(i:f)	30	1.38	.005	$.002<\sigma^2<.008$
ps(i:f)	6,480	.35	.353	$.343<\sigma^2<.363$

Table A-2. Estimated Variance Components for Personnel Specialist, Three-form Analysis

Effect	*Df*	*Ms*	σ^2	90% Confidence Intervals
Persons (p)	193	6.95	.047	$.024<\sigma^2<.069$
Sources (s)	2	146.98	.041	$.000<\sigma^2<.089$
Forms (f)	2	34.65	.002	$.000<\sigma^2<.014$
Items within f (i:f)	15	27.76	.045	$.018<\sigma^2<.072$
ps	386	3.74	.172	$.146<\sigma^2<.196$
pf	386	1.35	.023	$.013<\sigma^2<.033$
sf	4	1.87	.000	$.000<\sigma^2<.000$
psf	772	0.65	.043	$.034<\sigma^2<.052$
p(i:f)	2,895	0.68	.094	$.083<\sigma^2<.105$
s(i:f)	30	1.27	.005	$.002<\sigma^2<.008$
ps(i:f)	5,790	.40	.395	$.383<\sigma^2<.407$

Table A-3. Estimated Variance Components for Equipment Laboratory Specialist, Three-form Analysis

Effect	Df	Ms	σ^2	90% Confidence Intervals
Persons (p)	138	7.09	.087	$.055<\sigma^2<.119$
Sources (s)	2	22.37	.010	$.000<\sigma^2<.023$
Forms (f)	2	10.69	-.005	$.000<\sigma^2<.000$
Items within f (i:f)	12	21.34	.049	$.017<\sigma^2<.081$
ps	276	2.59	.140	$.116<\sigma^2<.164$
pf	276	1.08	.027	$.016<\sigma^2<.038$
sf	4	.23	-.001	$.000<\sigma^2<.000$
psf	552	.49	.033	$.023<\sigma^2<.043$
p(i:f)	1,656	.52	.065	$.054<\sigma^2<.076$
s(i:f)	24	.61	.002	$.000<\sigma^2<.004$
ps(i:f)	3,312	.32	.322	$.309<\sigma^2<.341$

Table A-4. Estimated Variance Components for Aerospace Ground Equipment, Three-form Analysis

Effect	Df	Ms	$\underline{\sigma}^2$	90% Confidence Intervals
Persons (p)	264	14.05	.122	$.093<\underline{\sigma}^2<.151$
Sources (s)	2	107.60	.016	$.000<\underline{\sigma}^2<.036$
Forms (f)	2	5.62	-.006	$.000<\underline{\sigma}^2<.000$
Items within f (i:f)	21	44.99	.054	$.027<\underline{\sigma}^2<.081$
ps	528	4.58	.160	$.141<\underline{\sigma}^2<.179$
pf	528	1.43	.022	$.016<\underline{\sigma}^2<.029$
sf	4	2.86	.000	$.000<\underline{\sigma}^2<.000$
psf	1,056	.74	.048	$.041<\underline{\sigma}^2<.055$
p(i:f)	5,544	.52	.055	$.049<\underline{\sigma}^2<.061$
s(i:f)	42	2.21	.007	$.004<\underline{\sigma}^2<.010$
ps(i:f)	11,088	.36	.359	$.351<\underline{\sigma}^2<.367$

Table A-5. Simulated D Study Results of Ratings Analysis for Aircrew Life Support

$\underline{\sigma}^2$ for pm(i:t) Design		$\underline{\sigma}^2$ for pM(I:T) Design				
	$\underline{n}_r$	1	1	1	3	3
	$\underline{n}_f$	1	2	4	1	4
	$\underline{n}_i$	8	4	8	16	8
$\underline{\sigma}^2{}_p$=.0884	$\underline{\sigma}^2{}_p$=	.0884	.0884	.0884	.0884	.0884
$\underline{\sigma}^2{}_s$=.0097	$\underline{\sigma}^2{}_S$=	.0097	.0097	.0097	.0032	.0032
$\underline{\sigma}^2{}_f$=.0006	$\underline{\sigma}^2{}_F$=	.0006	.0003	.0002	.0006	.0002
$\underline{\sigma}^2{}_{i:f}$=.0392	$\underline{\sigma}^2{}_{I:F}$=	.0049	.0049	.0012	.0025	.0012
$\underline{\sigma}^2{}_{ps}$=.1931	$\underline{\sigma}^2{}_{pS}$=	.1931	.1931	.1931	.0644	.0644
$\underline{\sigma}^2{}_{pf}$=.0284	$\underline{\sigma}^2{}_{pF}$=	.0284	.0142	.0071	.0284	.0071
$\underline{\sigma}^2{}_{sf}$=.0000	$\underline{\sigma}^2{}_{SF}$=	.0000	.0000	.0000	.0000	.0000
$\underline{\sigma}^2{}_{psf}$=.0613	$\underline{\sigma}^2{}_{pSF}$=	.0613	.0307	.0153	.0205	.0051
$\underline{\sigma}^2{}_{p(i:f)}$=.0736	$\underline{\sigma}^2{}_{p(I:F)}$=	.0092	.0092	.0023	.0046	.0023
$\underline{\sigma}^2{}_{s(i:f)}$=.0047	$\underline{\sigma}^2{}_{S(I:F)}$=	.0006	.0006	.0002	.0001	.0001
$\underline{\sigma}^2{}_{ps(i:f)}$=.3529	$\underline{\sigma}^2{}_{pS(I:F)}$=	.0441	.0441	.0110	.0074	.0037
	$\underline{\sigma}^2{}_p$=	.0884	.0884	.0884	.0884	.0884
	$\underline{\sigma}^2{}_\delta$=	.3362	.2913	.2289	.1252	.0826
	$\underline{\sigma}^2{}_\Delta$ =	.3520	.3068	.2401	.1316	.0872
	$\underline{\epsilon P}^2$ =	.208	.233	.279	.414	.517
	$\underline{\theta}$ =	.201	.224	.269	.402	.503

Table A-6. Simulated D Study Results of Ratings Analysis for Personnel Specialist

σ^2 for pm(i:t) Design	σ^2 for pM(I:T) Design					
	n_r	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	4	8	16	8
σ^2_p=.047	σ^2_p=	.0466	.0466	.0466	.0466	.0466
σ^2_s=.041	σ^2_S=	.0407	.0407	.0407	.0135	.0135
σ^2_f=.002	σ^2_F=	.0017	.0008	.0004	.0016	.0004
$\sigma^2_{i:f}$=.045	$\sigma^2_{I:F}$=	.0056	.0056	.0028	.0028	.0014
σ^2_{ps}=.172	σ^2_{pS}=	.1715	.1715	.1715	.0571	.0571
σ^2_{pf}=.023	σ^2_{pF}=	.0231	.0116	.0058	.0231	.0058
σ^2_{sf}=.000	σ^2_{SF}=	.0003	.0001	.0001	.0001	.0000
σ^2_{psf}=.043	σ^2_{pSF}=	.0426	.0213	.0107	.0142	.0036
$\sigma^2_{p(i:f)}$=.094	$\sigma^2_{p(I:F)}$=	.0118	.0118	.0030	.0059	.0030
$\sigma^2_{s(i:f)}$=.005	$\sigma^2_{S(I:F)}$=	.0006	.0006	.0001	.0001	.0001
$\sigma^2_{ps(i:f)}$=.395	$\sigma^2_{pS(I:F)}$=	.0494	.0494	.0124	.0082	.0041
	σ^2_p=	.0466	.0466	.0466	.0466	.0466
	σ^2_δ=	.2985	.2656	.2033	.1086	.0736
	σ^2_Δ =	.3473	.3135	.2460	.1269	.0893
	ϵP^2 =	.135	.149	.187	.300	.388
	θ =	.118	.129	.159	.269	.344

Table A-7. Simulated D Study Results of Ratings Analysis for Precision Measurement Equipment Laboratory Specialist

σ^2 for pm(i:t) Design		σ^2 for pM(I:T) Design				
	n_r	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	4	8	16	8
σ^2_{p}=.0868	σ^2_{p}=	.0868	.0868	.0868	.0868	.0868
σ^2_{s}=.0094	σ^2_{S}=	.0094	.0094	.0094	.0031	.0031
σ^2_{f}=.0000	σ^2_{F}=	.0000	.0000	.0000	.0000	.0000
$\sigma^2_{i:f}$=.0492	$\sigma^2_{I:F}$=	.0062	.0062	.0015	.0031	.0015
σ^2_{ps}=.1400	σ^2_{pS}=	.1400	.1400	.1400	.0467	.0467
σ^2_{pf}=.0265	σ^2_{pF}=	.0265	.0133	.0066	.0265	.0066
σ^2_{sf}=.0000	σ^2_{SF}=	.0000	.0000	.0000	.0000	.0000
σ^2_{psf}=.0334	σ^2_{pSF}=	.0334	.0167	.0083	.0111	.0028
$\sigma^2_{p(i:f)}$=.0648	$\sigma^2_{p(I:F)}$=	.0081	.0081	.0020	.0041	.0020
$\sigma^2_{s(i:f)}$=.0021	$\sigma^2_{S(I:F)}$=	.0003	.0003	.0001	.0021	.0000
$\sigma^2_{ps(i:f)}$=.3217	$\sigma^2_{pS(I:F)}$=	.0402	.0402	.0101	.0067	.0034
	σ^2_{p}=	.0868	.0868	.0868	.0868	.0868
	σ^2_{δ}=	.2483	.2183	.1671	.0951	.0615
	σ^2_{Δ} =	.2640	.2341	.1781	.1013	.0662
	ϵP^2 =	.259	.285	.342	.477	.585
	θ =	.248	.271	.328	.461	.568

Table A-8. Simulated D Study Results of Ratings Analysis for Aerospace Ground Equipment

$\underline{\sigma}^2$ for pm(i:t) Design		$\underline{\sigma}^2$ for pM(I:T) Design				
	$\underline{n}_r$	1	1	1	3	3
	$\underline{n}_f$	1	2	4	1	4
	$\underline{n}_i$	8	4	8	16	8
$\underline{\sigma}^2{}_p$=.1219	$\underline{\sigma}^2{}_p$=	.1219	.1219	.1219	.1219	.1219
$\underline{\sigma}^2{}_s$=.0159	$\underline{\sigma}^2{}_S$=	.0159	.0159	.0159	.0053	.0053
$\underline{\sigma}^2{}_f$=.0000	$\underline{\sigma}^2{}_F$=	.0000	.0000	.0000	.0000	.0000
$\underline{\sigma}^2{}_{i:f}$=.0536	$\underline{\sigma}^2{}_{I:F}$=	.0067	.0067	.0017	.0034	.0017
$\underline{\sigma}^2{}_{ps}$=.1599	$\underline{\sigma}^2{}_{pS}$=	.1599	.1599	.1599	.0533	.0533
$\underline{\sigma}^2{}_{pf}$=.0221	$\underline{\sigma}^2{}_{pF}$=	.0221	.0110	.0055	.0221	.0055
$\underline{\sigma}^2{}_{sf}$=.0001	$\underline{\sigma}^2{}_{SF}$=	.0001	.0001	.0000	.0000	.0000
$\underline{\sigma}^2{}_{psf}$=.0476	$\underline{\sigma}^2{}_{pSF}$=	.0476	.0238	.0119	.0159	.0040
$\underline{\sigma}^2{}_{p(i:f)}$=.0551	$\underline{\sigma}^2{}_{p(I:F)}$=	.0069	.0069	.0017	.0035	.0017
$\underline{\sigma}^2{}_{s(i:f)}$=.0070	$\underline{\sigma}^2{}_{S(I:F)}$=	.0009	.0009	.0002	.0002	.0001
$\underline{\sigma}^2{}_{ps(i:f)}$=.3593	$\underline{\sigma}^2{}_{pS(I:F)}$=	.0449	.0449	.0112	.0075	.0037
	$\underline{\sigma}^2{}_p$=	.1219	.1219	.1219	.1219	.1219
	$\underline{\sigma}^2{}_\delta$=	.2814	.2466	.1903	.1022	.0683
	$\underline{\sigma}^2{}_\Delta$=	.3050	.2701	.2081	.1111	.0753
	$\underline{\epsilon P}^2$ =	.302	.331	.391	.544	.641
	$\underline{\theta}$ =	.286	.311	.369	.523	.618

Table A-9. G Study Results for Crossed Design Analysis of WTPT Scores, Personnel Specialists

Effect	Df	Ms	σ^2	90% Confidence Intervals
Persons (p)	196	.70	.019	$.015<\sigma^2<.023$
Method (m)	1	28.24	.003	$.000<\sigma^2<.016$
Tasks (t)	3	9.07	-.005	$.000<\sigma^2<.000$
Items within t (i:t)	12	3.08	-.003	$.000<\sigma^2<.000$
pm	196	.18	-.014	$.000<\sigma^2<.000$
pt	588	.29	-.014	$.000<\sigma^2<.000$
mt	3	17.51	.016	$.000<\sigma^2<.039$
pmt	588	.40	.078	$.068<\sigma^2<.088$
p(i:t)	2,352	.09	.000	$.000<\sigma^2<.000$
m(i:t)	12	4.24	.021	$.000<\sigma^2<.044$
pm(i:t)	2,352	.09	.094	$.090<\sigma^2<.099$

Table A-10. G Study Results for Crossed Design Analysis of WTPT Scores, Aerospace Ground Equipment

Effect	Df	Ms	$\underline{\sigma}^2$	90% Confidence Intervals
Persons (p)	124	1.50	.006	$.004<\underline{\sigma}^2<.008$
Method (m)	1	18.[illegible]4	-.001	$.000<\underline{\sigma}^2<.000$
Tasks (t)	8	48.85	.004	$.002<\underline{\sigma}^2<.013$
Items within t (i:t)	90	5.89	-.008	$.000<\underline{\sigma}^2<.000$
pm	124	.38	.001	$.000<\underline{\sigma}^2<.000$
pt	992	.39	.001	$.000<\underline{\sigma}^2<.004$
mt	8	38.92	.022	$.012<\underline{\sigma}^2<.065$
pmt	992	.37	.020	$.018<\underline{\sigma}^2<.023$
p(i:t)	11,160	.14	-.004	$.000<\underline{\sigma}^2<.000$
m(i:t)	90	8.08	.053	$.050<\underline{\sigma}^2<.083$
pm(i:t)	11,160	.15	.149	$.146<\underline{\sigma}^2<.153$

Table A-11. G Study Results for Nested Design Analysis of WTPT Scores, Aircrew Life Support

Effect	Df	Ms	σ^2	90% Confidence Intervals
Persons (p)	192	1.14	.018	$.014<\sigma^2<.022$
Methods (m)	1	72.60	.004	$.000<\sigma^2<.021$
Tasks within m (t:m)	6	47.62	.026	$.001<\sigma^2<.051$
Items within t within m (i:t:m)	56	7.35	.037	$.025<\sigma^2<.049$
pm	192	.32	-.001	$.000<\sigma^2<.000$
p(t:m)	1,252	.33	.027	$.024<\sigma^2<.030$
p(i:t:m)	10,752	.12	.119	$.116<\sigma^2<.122$

Table A-12. G Study Results for Nested Design Analysis of WTPT Scores, Personnel Specialist

Effect	Df	Ms	σ^2	90% Confidence Intervals
Persons (p)	196	.63	.038	$.031<\sigma^2<.045$
Methods (m)	1	.01	-.007	$.000<\sigma^2<.000$
Tasks within m (t:m)	2	11.65	.013	$.000<\sigma^2<.022$
Items within t within m (i:t:m)	12	1.55	.008	$.004<\sigma^2<.012$
pm	196	.03	-.031	$.000<\sigma^2<.000$
p(t:m)	392	.28	.051	$.043<\sigma^2<.059$
p(i:t:m)	2,352	.07	.078	$.000<\sigma^2<.000$

Table A-13. G Study Results for Nested Design Analysis of WTPT Scores, Precision Measurement Equipment Laboratory Specialists

Effect	Df	Ms	σ^2	90% Confidence Intervals
Persons (p)	137	.37	.004	$.000<\sigma^2<.017$
Methods (m)	1	30.44	.004	$.000<\sigma^2<.015$
Tasks within m (t:m)	6	15.03	.010	$.000<\sigma^2<.023$
Items within t within m (i:t:m)	48	5.19	.037	$.025<\sigma^2<.049$
pm	137	.14	-.001	$.000<\sigma^2<.000$
p(t:m)	822	.17	.011	$.009<\sigma^2<.013$
p(i:t:m)	6,576	.09	.095	$.092<\sigma^2<.098$

Table A-14. G Study Results for Nested Design Analysis of WTPT Scores Aerospace Ground Equipment

Effect	*Df*	*Ms*	$\underline{\sigma}^2$	90% Confidence Intervals
Persons (p)	259	1.05	.011	$.009<\underline{\sigma}^2<.013$
Methods (m)	1	32.31	-.006	$.000<\underline{\sigma}^2<.000$
Tasks within m (t:m)	8	88 75	.036	$.002<\underline{\sigma}^2<.070$
Items within t within m (i:t:m)	70	13 92	.053	$.037<\underline{\sigma}^2<.069$
pm	259	18	-.006	$.000<\underline{\sigma}^2<.000$
p(t:m)	2,072	.42	.037	$.021<\underline{\sigma}^2<.053$
p(i:t:m)	18,130	.13	.126	$.123<\underline{\sigma}^2<.129$

Table A-15. Simulated D Study Results for WTPT Analysis of Personnel Specialist

σ^2 for pm(i:t) Design	σ^2 for pM(I:T) Design					
	n_m	1	1	1	2	2
	n_t	5	10	10	5	15
	n_i	5	5	15	15	10
σ^2_p=.0197	σ^2_p=	.0197	.0197	.0197	.0197	.0197
σ^2_m=.0035	σ^2_M=	.0035	.0035	.0035	.0017	.0017
σ^2_t=.0000	σ^2_T=	.0000	.0000	.0000	.0000	.0000
$\sigma^2_{i:t}$=.0000	$\sigma^2_{I:T}$=	.0000	.0000	.0000	.0000	.0000
σ^2_{pm}=.0000	σ^2_{pM}=	.0000	.0000	.0000	.0000	.0000
σ^2_{pt}=.0000	σ^2_{pT}=	.0000	.0000	.0000	.0000	.0000
σ^2_{mt}=.0165	σ^2_{MT}=	.0033	.0016	.0016	.0016	.0006
σ^2_{pmt}=.0780	σ^2_{pMT}=	.0156	.0078	.0078	.0078	.0026
$\sigma^2_{p(i:t)}$=.0000	$\sigma^2_{p(I:T)}$=	0000	.0000	.0000	.0000	.0000
$\sigma^2_{m(i:t)}$=.0211	$\sigma^2_{M(I:T)}$=	.0008	.0004	.0001	.0001	.0001
$\sigma^2_{pm(i:t)}$=.0938	$\sigma^2_{pM(I:T)}$=	.0038	.0019	.0006	.0006	.0003
	σ^2_p=	.0197	.0197	.0197	.0197	.0197
	σ^2_δ=	.0175	.0097	.0084	.0084	.0029
	σ^2_Δ =	.0247	.0152	.0137	.0120	.0053
	$E\rho^2$ =	.530	.670	.700	.700	.871
	θ =	.444	.564	.590	.622	.789

Table A-16. Simulated D Study Results for WTPT

Analysis of Aerospace Ground Equipment

$\underline{\sigma}^2$ for pm(i:t) Design		$\underline{\sigma}^2$ for pM(I:T) Design				
	$\underline{n}'_m$	1	1	1	2	2
	$\underline{n}'_t$	5	10	10	5	15
	$\underline{n}'_i$	5	5	15	15	10
$\underline{\sigma}^2_p$=.0055	$\underline{\sigma}^2_p$=	.0055	.0055	.0055	.0055	.0055
$\underline{\sigma}^2_m$=.0000	$\underline{\sigma}^2_M$=	.0000	.0000	.0000	.0000	.0000
$\underline{\sigma}^2_t$=.0044	$\underline{\sigma}^2_T$=	.0009	.0004	.0004	.0009	.0003
$\underline{\sigma}^2_{i:t}$=.0000	$\underline{\sigma}^2_{I:T}$=	.0000	.0000	.0000	.0000	.0000
$\underline{\sigma}^2_{pm}$=.0001	$\underline{\sigma}^2_{pM}$=	.0001	.0001	.0001	.0001	.0001
$\underline{\sigma}^2_{pt}$=.0013	$\underline{\sigma}^2_{pT}$=	.0003	.0001	.0001	.0003	.0001
$\underline{\sigma}^2_{mt}$=.0223	$\underline{\sigma}^2_{MT}$=	.0045	.0022	.0022	.0022	.0007
$\underline{\sigma}^2_{pmt}$=.0202	$\underline{\sigma}^2_{pMT}$=	.0041	.0020	.0020	.0020	.0007
$\underline{\sigma}^2_{p(i:t)}$=.0000	$\underline{\sigma}^2_{p(I:T)}$=	.0000	.0000	.0000	.0000	.0000
$\underline{\sigma}^2_{m(i:t)}$=.0634	$\underline{\sigma}^2_{M(I:T)}$=	.0025	.0013	.0004	.0004	.0002
$\underline{\sigma}^2_{pm(i:t)}$=.1489	$\underline{\sigma}^2_{pM(I:T)}$=	.0060	.0030	.0010	.0010	.0005
	$\underline{\sigma}^2_p$=	.0055	.0055	.0055	.0055	.0055
	$\underline{\sigma}^2_\delta$=	.0104	.0053	.0033	.0034	.0013
	$\underline{\sigma}^2$ =	.0183	.0092	.0064	.0069	.0026
	$\underline{\epsilon P}^2_\Delta$ =	.348	.513	.628	.624	.807
	$\underline{\theta}$ =	.232	.376	.466	.447	.683

SUPPLEMENTARY

INFORMATION

ERRATA

90-21

AD-A225 011

AIR FORCE HUMAN RESOURCES LABORATORY
BROOKS AIR FORCE BASE, TEXAS 78235-5601

ERRATA

Kraiger, K. (1990, July). Generalizability of Walk-Through Performance Tests, Job Proficiency Ratings, and Job Knowledge Tests Across Eight Air Force Specialties (AFHRL-TP-90-14, AD-A225 011). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

A corrected page 31 is attached to replace the one printed in the original technical paper.

ESTHER M. BARLOW
Technical Editing

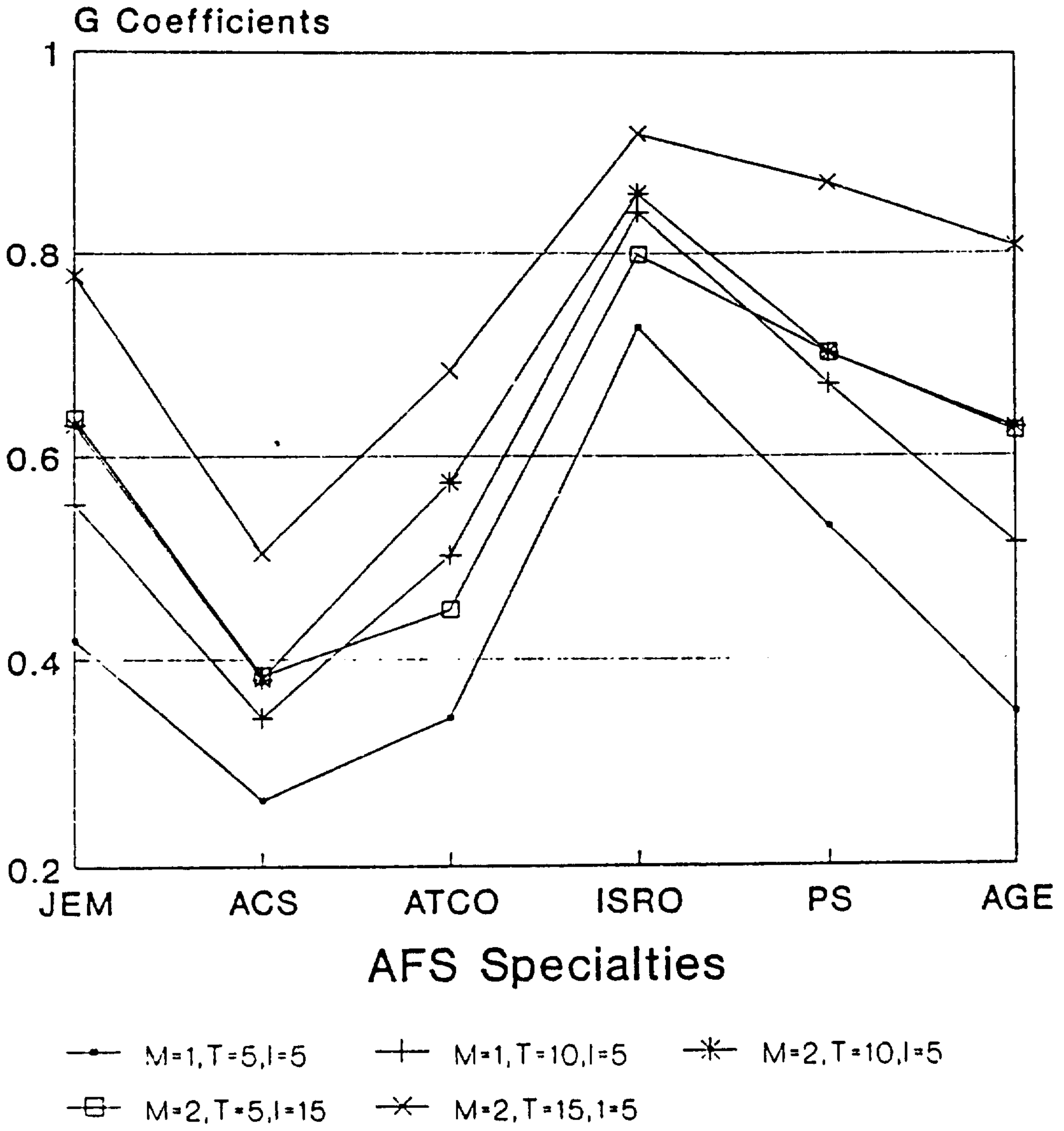

Figure 3. G Coefficients for WTPT Scores for Six Occupational Specialties